HIV-1(V3 loop)とエントロピー (符号と暗号の代数的数理)

AUTHOR(S):

多田, 秀樹; 関田, 英太郎

CITATION:

多田, 秀樹 ...[et al]. HIV-1(V3 loop)とエントロピー (符号と暗号の代数的数理). 数理解析研究所講究録 2005, 1420: 18-27

ISSUE DATE:

2005-04

URL:

http://hdl.handle.net/2433/47178

数理解析研究所講究録 1420 巻 2005 年 18-27

HIV-1(V3 loop) とエントロピー

多田 秀樹　　関田英太郎
Hideki TADA　Eitarou SEKITA

法政大学大学院工学研究科システム工学専攻
日本獣医畜産大学

Abstract The V3 loop is a region in the envelope protein of HIV-1 whose polypeptide(amino acid sequence) extremely tends to change.Certain amino -acid pairs in the V3 loop do not have only independent mutations but covariant mutations.We apply information theory for an analysis of the V3loop in order to draw a certain law of change of polypeptide of the V3 loop.

Key Words : V3 loop,Mutual Information,Specific Information,Covariant Mutation flow

1 はじめに

現在、世界的に猛威を振るっているウィルスの一つに HIV ウィルスがある。HIV-1 の起源は、霊長類を自然宿主とするサル免疫不全ウィルス (Siman immunodeficiency virus,SIV) の人への伝播によるという有力な証拠が集まりつつある。WHO の報告では、2002 年 12 月現時点での世界の HIV 感染者総数は 4200 万人、年間新規感染者数は約 500 万人と報告されている。世界の総人口 (約 62 億) の約 150 人に一人が感染しており、1 日当り 13700 人、実に 6 秒当り 1 人の新たな感染者が発生している計算となる。アフリカのいくつかの国々では AIDS の流行により平均寿命が 60 歳から 40 歳にまで減少し、深刻な社会問題となっている。HIV 感染予防の究極の方法はワクチンの開発である。しかしながら、HIV ウィルスが抗原構造の多様性と著しい変異性を持つこと、HIV が免疫応答の中枢にあるヘルパー T 細胞そのものを破壊してしまうこと、さらにはワクチン開発のための優れた動物モデルが無いなどのさまざまな理由からワクチンの実用化の目処は未だ立たないのが現状である。

今回の実験では、HIV-1 ウィルスのアミノ酸配列の中の特に V3loop という箇所を実験の対象として、アミノ酸配列解析を行う。V3loop は HIV-1 のアミノ酸配列の中でも特に劇的な変異性を持つ箇所として知られ、このことが HIV に有効なワクチン等の開発の妨げになるとも考えられる。アミノ酸の変異を考えた時、ループ内のアミノ酸では、幾つかのアミノ酸どうしが何らかの連携を図って変化 (共変) している可能性も伺える。そこで、情報理論の考えを用い、多変数相互情報量を共変値としてとらえる事で、V3loop 内のアミノ酸の相互依存性を探り、アミノ酸間に存在する可能性のあるアミノ酸ネットワークを描くことを目的とした。

2 基本事項

2.1 タンパク質

多くの生物の遺伝情報は、4種類の塩基(アデニン、チミン、シトシン、グアニン)からなるDNAという形で蓄えられている。DNAは、転写をおこしRNAとなり、これが翻訳されることによって必要なタンパク質が合成されている。しかし、今回取り扱うHIVウィルスは、生物界で唯一逆転酵素をもつレトロウィルスと呼ばれるものに属し、遺伝情報はRNAの形で蓄えられている。タンパク質は、一本の鎖状につながった20種類のアミノ酸が複雑な立体構造を成して形成されていて、タンパク質ごとに構成アミノ酸の種類・数・結合順は異なり、それによって様々な機能をもつ分子となっている。この構造及び機能を解析することはポストゲノムと呼ばれ、多くの研究機関が取り組んでいる問題でもある。以下に20種類のアミノ酸とそのアミノ酸コードを示す。

表1 アミノ酸コード表

A	*Alanine*	*R*	*Arginine*
N	*Asparagine*	*D*	*Asparticacid*
C	*Cysteine*	*Q*	*Glutamine*
E	*Glutamicacid*	*G*	*Glycine*
H	*Histidine*	*I*	*Isoleucine*
L	*Leucine*	*K*	*Lysine*
M	*Methionine*	*F*	*Phenylalanine*
P	*Proline*	*S*	*Serine*
T	*Threonine*	*W*	*Tryptophan*
Y	*Tyrosine*	*V*	*Valine*

2.2 HIVウィルスとV3loop

HIVは直径110nm、約9500塩基からなるRNA型エンベロープウィルスである。逆転酵素などを含むキャプシドと、それを取り囲むエンベロープにより構成されている。ウィルス粒子の外側を構成するエンベロープには、糖タンパク質gp120と糖タンパク質gp41があり、gp120内のV3loop領域は非常に変化の富んだ領域(様々な配列が現れる)であり、機能上・免疫学上の両面において重要とされている。

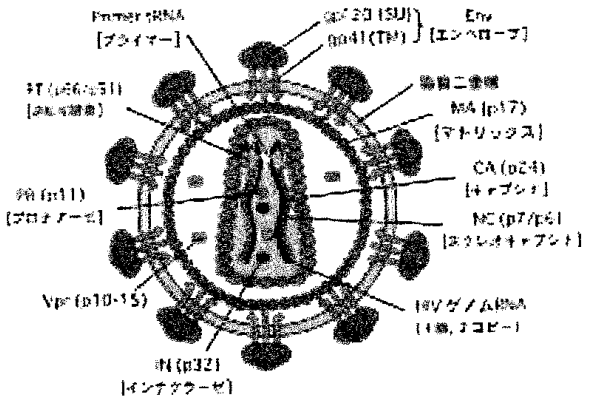

図1 HIV

V3loop 領域は、両末端をシステイン (C) でジスルフィド結合をしたループ構造をしていて、抗体や T 細胞を中和する効力のある抗原エピトープである。 図 2 は V3loop の一例の抽象的なモデル図である。アルファベットはそれぞれアミノ酸コードに対応し、それらの右下に添えてある数字 (i) は解析を行う上での便宜上の位置 (site) を表している。また、V3loop の先端にあたるアミノ酸配列が GPGQ であればアフリカ分離株、GPGR であれば欧米分離株と呼ばれる (注：異なる配列モデルの場合、アミノ酸の配列は異なる)。

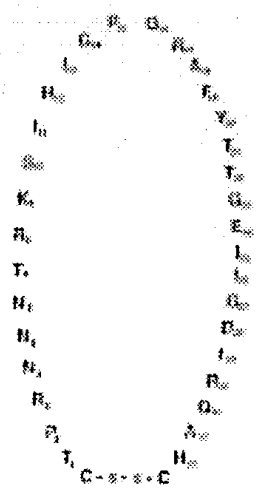

図 2 V3loop のモデル図

図 2 の各位置 (i) における、20 種のアミノ酸が現れる確率 (生起確率) を、データベース (National Center for Biotechnology Information(NCBI)[7]) に実在する約 17000 例の V3 ループの配列から求め、各位置における生起確率を用いるて、アミノ酸の変化量をエントロピーの概念を用いて定量的に表現する。

また本論文では、異なる位置にあるアミノ酸が、なんらかの関連をもって同時に変化することを共変と呼ぶことにし、多変数相互情報量を用いることで、V3loop 内に存在するであろう共変グループの特定を目的とする。

2.3 アライメント (alignment)

アライメントとは、配列の類似する部分、または同一となる部分を縦に揃えて並べ合わせる操作のことを言い、アミノ酸 (塩基) 配列の類似性解析の基本的なものの一つである。二本のアミノ酸配列 HEAGAWGHEE と PAWHEAE のアラインメントを行うと以下のようになる。

H E A G A W G H E E
P A W H E A E
⇓
H E A G A W G H E - E
- - P - A W - H E A E

図 3 アライメント

同じアミノ酸や性質の似たアミノ酸が縦に、同じカラム位置になるように、ところどころハイフン (−) が挿入されている。生化学的には gap(ギャップ) は遺伝子の欠損等によりア

ミノ酸が出現していないことを意味する。縦方向に揃った代表的文字で構成される配列をコンセンサス配列 (consensus sequence) と呼び、この中に見られるパターンがアライメントされた配列群を特徴付けるものと判断できるとき、このパターンをモチーフ (motif) と呼ぶ。また、gap の入り方の少ない領域を保存部位 (conservative site) と言い、保存性の高い部位は配列の中でもタンパク質の構造や機能の実現の上で重要な部分であると推測できる。重要な配列部分は進化の過程の中で保存されているため、生き残っている生物の同種のタンパク質をアライメントで調べると保存部位を発見できる。

アライメントの内、対象となる配列が 2 本のものをペアワイズアライメント (pairwise alignment) と呼び、基本的には配列と配列の類似性を求めることに使われる。また、3 本以上の配列を扱う場合はマルチプルアライメント (multiple alignment) と呼んで区別され、ペアワイズアライメントよりも効果的に配列の共通性を見出せる。

2.4 エントロピーと多変数相互情報量

V3loop のそれぞれの位置では様々なアミノ酸が現れる場合もあれば、逆にほぼ決まったアミノ酸が現れる場合もある。そこで V3loop の位置 $i(1 \leqq i \leqq 33)$ に現れるアミノ酸を A^i で表す。ここで、A は 20 種類のアミノ酸コードに対応するアルファベット 1 文字が表記される。あるアミノ酸の生起確率を $P(A^i)$ として、位置 i においてどの程度アミノ酸が変化するか (不確実度) を一般にエントロピー

$$H(i) = -\sum_{A^i} P(A^i) \log P(A^i)$$

で表す。エントロピー H(i) は非負である。また、現れるアミノ酸が多様であるほど高い値を示す。ここで、位置 i に A^i, 位置 j に A^j が現れる生起確率を $P(A^i, A^j)$ $(i \neq j)$ とすると

$$H(i,j) = -\sum P(A^i, A^j) \log P(A^i, A^j)$$

を結合エントロピーと呼ぶ。

アミノ酸が現れる位置 i と j をペアとして捉えることで、両位置間の共変性をあらわす事ができる。その際、共変性の尺度として相互情報量を用いることができる。相互情報量は $M(i,j)$ は非負である。

$$\begin{aligned} M(i,j) &= H(i) - H(i|j) \\ &= H(j) - H(j|i) \\ &= H(i) + H(j) - H(i,j). \end{aligned}$$

ただし、$H(i|j)$ は条件付エントロピーとする。

$$H(i|j) = -\sum P(A^i) P(A^i|A^j) \log P(A^i|A^j).$$

ここで、2 位置 i,j におけるアミノ酸が完全な共変動をするとき、$M(i,j)$ は最大値となる。最小値 0 になるときは、位置 i,j が完全に独立な変化をするか、変化がないときに達する。つまり、アミノ酸ペアの多様性が大きければ $M(i,j)$ の値は大きくなる。

3 多位置間相互情報量

本論文での目的は、相互情報量を用いた 2 位置間における共変の数値化 [1] を拡張し、多変数相互情報量を用いることで共変グループの特定することである。多位置間におけ

るの多変数相互情報量は負の値も取り得る。負の相互情報量とは情報理論の観点からすると、異なるいくつかの情報源から得られる情報により、情報を得た側のエントロピーが増大したときにおこりうる。これを共変という観点から見ると、負の相互情報量が現れる位置間では共変を阻害するように働く位置が含まれていることを意味する。仮に位置 $(1 \leq j \leq m-1)$ の任意の位置間での多位置間相互情報量が全て正の値をとるとき、位置 m を加えた m 位置間での多変数相互情報量ではじめて負の値をとったとすると、この位置 m は共変のグループに含まれていないと考えることができる。

多位置間 $(1 \leq j \leq n)$ の多変数相互情報量は、$(1 \leq j \leq n-1)$ 位置間の相互情報量と条件付き相互情報量を用いて以下の様に表される。

$$M(i_1, \cdots, i_n) = M(i_1, \cdots, i_{n-1}) - M(i_1, \cdots, i_{n-1}|i_n). \quad \cdots(1)$$

位置 i_n にアミノ酸 $A^{(n)}$ が現われたとき位置 i_j $(1 \leq j \leq n)$ にアミノ酸 $A^{(j)}$ の現われる条件付確率を $P(A^{(1)}, \cdots, A^{(n-1)}|A^{(n)})$ と記すとき、条件つきエントロピーは

$$H(i_1, i_2, \cdots, i_{n-1}|i_n)$$
$$\stackrel{def}{=} -\sum_{A^{(j)}} P(A^{(n)}) P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) \log P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}).$$

右辺の和は $1 \leq j \leq n$ について各々 20 種類のアミノ酸 $A^{(j)}$ を渡る。

$1 \leq j \leq n-1$ について位置 i_j にアミノ酸 $A^{(j)}$ が現われる結合事象の生じる確率は

$$P(A^{(1)}, A^{(2)}, \cdots, A^{(n)}) = P(A^{(n)}) \cdot P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) \quad \cdots(2)$$

であり、結合エントロピーは

$$H(i_1, i_2, \cdots, i_n) \stackrel{def}{=} -\sum_{A^{(j)}} P(A^{(1)}, A^{(2)}, \cdots, A^{(n)}) \log P(A^{(1)}, A^{(2)}, \cdots, A^{(n)})$$

と一般に表される。

Prop. $H(i_1, i_2, \cdots, i_{n-1}|i_n) = H(i_1, i_2, \cdots, i_n) - H(i_n)$

(proof) (2) より

$$\begin{aligned}
H(i_1, i_2, \cdots, i_n) &= -\sum_{A^{(j)}} P(A^{(1)}, A^{(2)}, \cdots, A^{(n)}) \log P(A^{(1)}, A^{(2)}, \cdots, A^{(n)}) \\
&= -\sum_{A^{(j)}} P(A^{(n)}) P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) \\
&\qquad \left\{ \log P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) + \log P(A^{(n)}) \right\} \\
&= -\sum_{A^{(j)}} P(A^{(n)}) P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) \\
&\qquad \log P(A^{(1)}, A^{(2)}, \cdots, A^{(n-1)}|A^{(n)}) \\
&\quad - \sum_{A^{(n)}} P(A^{(n)}) \log P(A^{(n)}) \left\{ \sum_{A^{(1)}} \cdots \sum_{A^{(n-1)}} P(A^{(1)}, \cdots, A^{(n-1)}) \right\} \\
&= H(i_1, i_2, \cdots, i_{n-1}|i_n) + H(i_n). \quad \blacksquare
\end{aligned}$$

$n=2$ のときの相互情報量

$$M(i_1,i_2) = H(i_1) - H(i_1|i_2)$$
$$= -\sum_{A^{(1)}} P(A^{(1)}) \log P(A^{(1)}) + \sum_{A^{(1)},A^{(2)}} P(A^{(2)})P(A^{(1)}|A^{(2)}) \log P(A^{(1)}|A^{(2)}) \quad \cdots (3)$$

は良く知られたものである。われわれはこの関係式を一般化して、$n \geq 3$ に対しても相互情報量及び条件付相互情報量を考える;

$n=3$ のとき

$$M(i_1,i_2,i_3) = M(i_1,i_2) - M(i_1,i_2|i_3).$$

右辺の $M(i_1,i_2|i_3)$ は条件付相互情報量であって

$$M(i_1,i_2|i_3) \stackrel{def}{=} -\sum_{A^{(1)},A^{(3)}} P(A^{(3)})P(A^{(1)}|A^{(3)}) \log P(A^{(1)}|A^{(3)})$$
$$+ \sum_{A^{(1)},A^{(2)},A^{(3)}} P(A^{(3)})P(A^{(2)}|A^{(3)})P(A^{(1)}|A^{(2)}|A^{(3)}) \log P(A^{(1)}|A^{(2)}|A^{(3)}).$$

これは (3) を条件付確率で置き換えることにより定義される。

$P(A^{(1)}|A^{(2)}|\cdots|A^{(n)}) = \dfrac{P(A^{(1)},A^{(2)},\cdots,A^{(n)})}{P(A^{(2)})\cdots P(A^{(n)})}$ より

$$M(i_1,i_2|i_3) = -\sum P(A^{(1)},A^{(3)}) \log \frac{P(A^{(1)},A^{(3)})}{P(A^{(3)})}$$
$$+ \sum P(A^{(2)},A^{(3)}) \frac{P(A^{(1)},A^{(2)},A^{(3)})}{P(A^{(2)})P(A^{(3)})} \log \frac{P(A^{(1)},A^{(2)},A^{(3)})}{P(A^{(2)},A^{(3)})}$$
$$= -\sum_{A^{(1)},A^{(3)}} P(A^{(1)},A^{(3)}) \log P(A^{(1)},A^{(3)}) + \sum_{A^{(3)}} P(A^{(3)}) \log P(A^{(3)}) \sum_{A^{(1)}} P(A^{(1)}|A^{(3)})$$
$$+ \sum_{A^{(1)},A^{(2)},A^{(3)}} P(A^{(1)},A^{(2)},A^{(3)}) \log P(A^{(1)},A^{(2)},A^{(3)})$$
$$- \sum_{A^{(2)},A^{(3)}} P(A^{(2)},A^{(3)}) \log P(A^{(2)},A^{(3)}) \sum_{A^{(1)}} P(A^{(1)}|(A^2,A^{(3)})).$$

ここで $\sum\limits_{A^{(1)}} P(A^{(1)}|A^{(3)}) = \sum\limits_{A^{(1)}} P(A^{(1)}|(A^2,A^{(3)})) = 1$ だから

$$M(i_1,i_2|i_3) = H(i_1,i_3) - H(i_3) - H(i_1,i_2,i_3) + H(i_2,i_3)$$

が得られる。

Prop. $M(i_1,i_2,i_3) = \sum\limits_{j=1}^{3} H(i_j) - \sum\limits_{1\leq j<k\leq 3} H(i_j,i_k) + H(i_1,i_2,i_3).$

(proof) $M(i_1,i_2,i_3) = M(i_1,i_2) - M(i_1,i_2|i_3)$ に

$$M(i_1,i_2) = H(i_1) + H(i_2) - H(i_1,i_2),$$
$$M(i_1,i_2|i_3) = H(i_1,i_3) - H(i_3) - H(i_1,i_2,i_3) + H(i_2,i_3)$$

を代入すればよい。 ■

同様に、$M(i_1,i_2,\cdots,i_n)$ は $M(i_1,\cdots,i_{n-1})$ と条件付確率を考えた $M(i_1,\cdots,i_{n-1}|i_n)$ より求められ、多変数相互情報は結合エントロピーを用いて一般に以下の様に表される。

$$M(i_1,\cdots,i_n)$$
$$=\sum_{j1=1}^{n}H(i_{j1})-\sum_{1\le j1<j2\le n}H(i_{j1},i_{j2})+\sum_{1\le j1<j2<j3\le n}H(i_{j1}.i_{j2},i_{j3})-\cdots$$
$$\cdots+(-1)^{n-2}\sum_{1\le j1<j2<\cdots<j(n-1)\le n}H(i_{j1},i_{j2},\cdots,i_{j(n-1)})+(-1)^{n-1}H(i_1,i_2,\cdots,i_n).$$

(証明略)

4 共変指数

各位置でのエントロピーをふまえた上で位置間の相互エントロピーを考えた時、、エントロピーの低い位置間よりも高い位置間の方が相互エントロピーの値は高い値をとってしまう。しかしながら、エントロピーの低い位置においても共変を起こしている可能性は十分に考えられ、相互エントロピーの値をそのままグループの特定に至る共変の強さを表す尺度として用いるのは妥当ではないと考えられる。そこで、次式を共変指数：K として定義する。

$$K(i_1,\cdots,i_n)=\frac{M(i_1,\cdots,i_n)}{n\sum_{i=1}^{n}H(i)}.$$

ここで、n 位置間未満での相互エントロピーは全て正の値を示すもののみを考える。これは 3 節でも前述したが、仮に位置 $(1\le j<m<n)$ の任意の位置間での j 位置間相互情報量が全て正の値をとり、位置 m を加えた m 位置間での多変数相互情報量ではじめて負の値をとったとすると、この位置 m は共変のグループに含まれていないと考えることができ、この時点で共変していない位置 m が含まれてしまうためである。

このように考えると、上式は、n 位置のエントロピーに対する、n 位置間相互エントロピーの平均の割合を示すこととなり、この値が 1 に近いほど n 位置間の共変の強さは高いものとして考えられる。さらに、実験より導くことが可能であろう共変値 k(n) を与えることで、共変グループの特定ができるものと考える。

5 データ群の作成

V3 領域に関するアミノ酸配列のデータは NCBI のホームページから約 17000 例取り出す事ができる。このさい、本論文では HIV-1 の V3loop のみを主眼においているため疑わしい配列は排除した。その方法として PAPIA システム (Parallel Protein Information Analysis System(PAPIA),http://www.cbrc.jp/papia-cgi/mul_queryJ.pl 参照) によるアライメントを用いて V3loop の正否を判定した。さらに、HIV-1 の V3loop におけるアミノ酸配列で位置 14〜17 付近が GPGR(欧米分離株) と GPGQ(アフリカ分離株) は別のものであると既に大別されているため、この 2 種にデータ群を GPGR(8340 例) と GPGQ(4033 例) に大別した。このようにして得られたデータを最新版 clustalW にかけ、マルチプルアライメントを行った。従来の clustalW ver1.82 には制限 (最長配列×配列数 $\le$ 10000) があるため、一度に何千本の配列のアライメントを行うことが不可能であったが、今回、国立遺伝学研究所の全面協力により実現した。

6 実験, 実験データ

各位置におけるエントロピーを計算することで、アミノ酸がどの程度変化しているのかを求める。また、2位置間・3位置間での相互情報量を算出する事により、アミノ酸間での共変性が高いと思われる位置を示す。今回の実験ではPCのスペック等の問題により、3位置間までのデータでしか実験を行えず、また、データの処理等にも間に合わなかったため、共変指数についても考察するにいたらなかった。しかしながら前述の手法を用いて遺伝領域全体までのデータを取ることで、アミノ酸間に存在するネットワークを求めることができると考えている。以下のグラフは図4を除いて、結果を一部掲載したものである。

6.1 位置 i におけるエントロピー

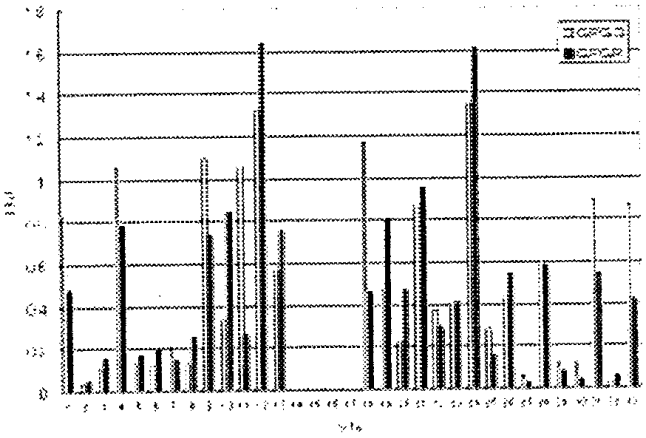

図4 エントロピー $H(i)$

位置 $14 \leq i \leq 17$ は、GPGR型,GPGQ型のみを実験の対象としているため、複雑性を示すエントロピーの値は0となる。

6.2 位置 (i, j) における相互情報量

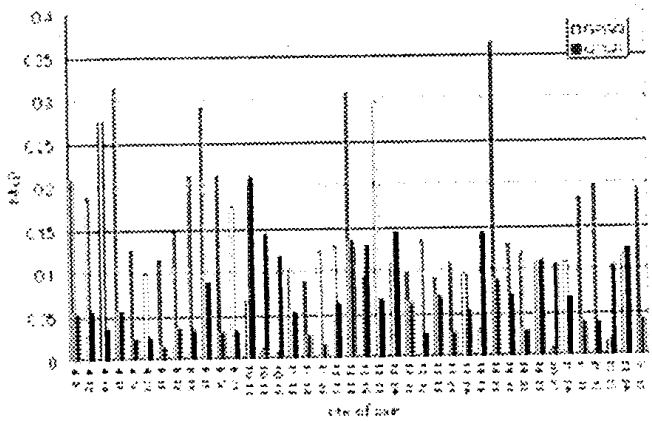

図5 相互情報量 $M(i, j)$

図5ではGPGQ型とGPGR型のどちらかの相互情報量が比較的高い値となった位置を、比較のために両型共に掲載した。

6.3 位置 (i,j,k) における相互情報量

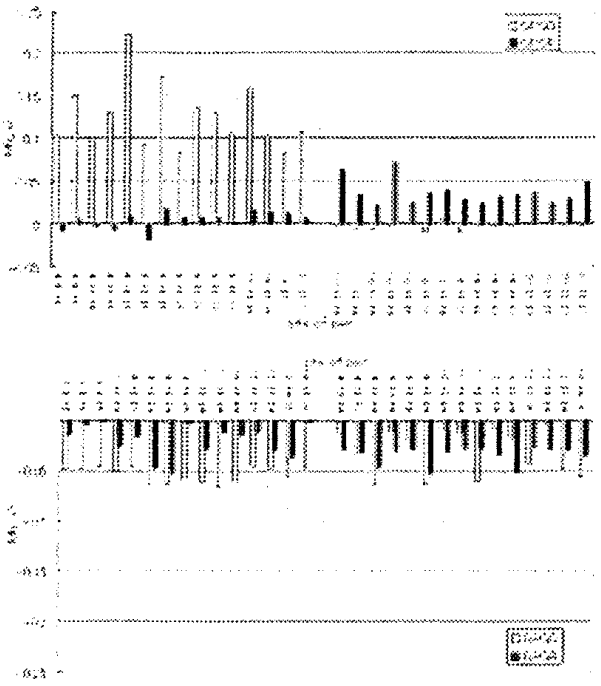

図 6 相互情報量 $M(i,j,k)$

7 考察

エントロピーについて図 4 より、両型共に位置 12 や 24 ではアミノ酸の保存性が低く、逆に位置 2,3,5,6,7,8,27,29,30,32 では保存性が高い。保存性の高い位置、及びそこに現れるアミノ酸は、全ての V3loop において共通してのものであり、HIV ウィルスが存続していく上で限りなく重要であり、かつ基本的な構造や機能に貢献している部分だと推測される。型別に見ると最大値を誇るのは GPGR 型であるが、広範囲にわたって高い値を示しているのは GPGQ 型であり、相互情報量も同様に GPGQ 型の方が全体的に高い値を取っている。このことより、GPGQ 型の方が平均的に変異性及び共変性が強い事を思わせる。これは、GPGQ 型がアフリカ株たる所以を数値的に読み取れたものと考えられる。なぜなら、HIV の患者はアフリカで爆発的に増加・拡大し、そしてその増加は頭打ちになりつつも現在進行形であるため、その過程において激しく変異が起こっていると容易に考えれられるからである。

多位置間相互エントロピーについて、図 5 で 2 位置間相互情報量が高い値をとる位置 4,18,21 の位置を注目すると、図 6 よりこの 3 位置間相互エントロピーにおいても、高い値を示しており、この 3 位置は強い関係性 (共変) を持つものと考えられる。しかし、やはり図 5 で高い値を示す位置 4,12,33 の位置に注目してみると、、図 5、6 よりこの 3 位置間相互エントロピーはマイナスの値をとり、任意の 2 位置間を考えた時には共変しているが、3 位置間でみると共変していないと確認できる。すなわちこの 3 位置は同じ共変のグループには属していないものと考えられる。V3loop の全ての位置に対し、多位置間相互エントロピーを考え、今回検証を行えなかったが、さらに共変指数を用いることでアミノ酸の間に存在するであろうアミノ酸ネットワークを明らかにすることができると考えている。

参考文献

[1] Bette T.M.Korber, Robert M.Farber, David H.Wolpert, Alan S.Lapedes, Covariation of mutation in the V3 loop of human immunodeficiency virus type 1 envelope protein, An information theoretic analysis, Proc.Natl.Acad.Sci.USA, Vol.**90**, pp.7176-7180, August 1993.

[2] Bruce Alberts et al. 中村桂子他訳, 細胞の分子生物学第 3 版,Newton Press,1995.

[3] Oya M.:"Information Theoretial Treatment of Genes",Trans.IEICE,E72,5,pp.556-560,May 1989.

[4] T.Tsujishita,On Triple Mutual Information,Adva
nces In Applied Math,16,pp269-274,1995.

[5] 今井秀樹, 情報理論, 昭晃堂,1984.

[6] 中山晃治, 田辺文雄, 多田秀樹, 関田英太郎, 平松豊一, 西村滋人, エイズウィルスにおける V3 ループの情報理論的解析, 第 25 回情報理論とその応用シンポジウム予稿集,pp763-766,2002.

[7] National Center for Biotechnology Information, ht
tp://www4.ncbi.nlm.nih.gov/Entrez/